AR 3261

Fidan Limani, Atif Latif, Timo Borst, and Klaus Tochtermann

Metadata Challenges for Long Tail Research Data Infrastructures

Abstract: Research Data has emerged as a 1st class research citizen, providing prospects for new services, projects, and ultimately users. In the context of long tail research data, we aim to make this data available to researchers across disciplines via a set of metadata-based services. The metadata component – GeRDI schema – represents one of the key challenges faced in this undertaking.

Keywords: Metadata schema, long tail research data, research data infrastructure

Herausforderungen mit Metadaten für Long-Tail-Forschungsdaten Infrastrukturen

Zusammenfassung: Forschungsdaten hat sich zu einem vordringlichen Thema entwickelt. Es bietet Perspektiven für neue Dienste, Projekte und letztlich Nutzer. Unser Ziel ist es, Forschern Long-Tail-Forschungsdaten über eine Reihe von metadatenbasierten Diensten zugänglich zu machen. Die Metadatenkomponente – das GeRDI Schema – stellt eine der zentralen Herausforderungen bei diesem Vorhaben dar.

Schlüsselwörter: Metadatenschema, Long-Tail-Forschungsdaten, Forschungsdaten-Infrastruktur

1 Introduction

The advent of data intensive research has fueled the creation and dissemination of research data (RD). However, there is still a need for new approaches to position it for general re-usability – support sharing, reproducibility, citation, etc., for different RD types. RD infrastructures (RDI) package RD and services, to lower the barrier for RD-centered activities. There are multiple such initiatives that provide solutions across domains, geographies, and scopes.

In Germany, policy discussions to frame existing and guide future RDI initiatives are underway. The Union of German Academies of Sciences and Humanities encourages and considers them as an urgent undertaking. Such infrastructures would enable exchange of RD and research practices, which translates to increased RD utilization within and across disciplines – creating potential research synergies – in an

otherwise isolated and underutilized RD scene.[1] In a similar tone, in a report about RDI approaches, the Council for Information Infrastructures (RfII) emphasizes the need for interoperability and coordination between different RD infrastructures in order to extend the current infrastructure reach, and eventually converge towards a national, multidisciplinary RDI that could serve a broader spectrum of research domains and users.[2] We see national and local RDI undertakings with similar objectives in EU, US, Canada, etc. See the "Related work" section for more.

The focus of RDI projects is usually on established, domain-specific research communities. One area usually left out of these projects is that of long tail RD. This RD typically consists of small RD "contributions" (low in volume), heterogeneous – spanning multiple research disciplines, with no established RD management practices, such as metadata standards to adhere to, RD storage and dissemination setup, etc.[3] In the midst of RDI project initiatives, we see a need to target communities that lack established RD management practices and associated infrastructures. Long tail RD metadata heterogeneity presents one of the key enablers and, at the same time, challenges for supporting such an RDI. In this work we propose an approach that maintains a reasonable balance between generic and disciplinary metadata to support key use cases across (and between) research communities in an RDI.

The paper is structured as follows: we provide context – scope and objectives – for project GeRDI,[4] a generic RDI that focuses on long tail RD. Next, we highlight the role of metadata as a key enabler for data discoverability, schema creation and further services development. We then discuss the GeRDI design, from its pilot community requirements, to metadata standard options and schema finalization. A brief review of implemented services and their relation to the metadata is then provided. We end the paper with review of related work and conclusions and plans for future work.

2 GeRDI Project: A Generic Research Data Infrastructure

With the rise of data-enabled research, researchers expect seamless, infrastructure-like support to reuse, reproduce, derive from, etc., existing and create new RD outcomes, thus RDIs present a key enabler for this scenario. GeRDI – the Generic RD Infrastructure – is a distributed and federated RDI that focuses on long tail RD.[5] The project aims to integrate research (meta)data from multiple communities and research disciplines, and provide users with a "one stop shop" in terms of research (meta)data, thus

[1] Union der Deutschen Akademien der Wissenschaften (2018).

[2] RfII (2017) 53.

[3] e-IRG (2016); Heidorn (2008).

[4] https://www.gerdi-project.eu.

[5] Grunzke et al. (2017).

eliminate the need to switch RDI platforms based on disciplines they support, datasets collections they contain, or services they offer. Moreover, the presence of multi-disciplinary RD in the same infrastructure could potentially trigger research synergies.

GeRDI adopts open standards and best practices to position itself well in the infrastructure context, and establishes sufficient alignment with both national and international projects, such as NFDI[6] in Germany, or EOSC[7] pilot at EU level. This is important for the interoperability desiderata of RDI projects in Europe and more. Such is the case with bringing RD metadata to a FAIR[8] level – to the extent possible – as one of the features of the project. GeRDI faces a variety of research domains, practices, standards, and services, from highly structured to unstructured research communities. From the metadata perspective, this implies dealing with multiple metadata standards and application practices across communities (different metadata granularity, coverage, etc.). To provide a glimpse of this (not only) RD variety, Table 1 lists its pilot communities, categorized by the disciplines they cover.

Research community	Research focus
Social sciences and Economics	- Socio-Economic Panel
Life sciences and Humanities	- Microscopy and Bioinformatics - Digital Humanities - National Center for Tumor Diseases
Marine sciences	- Environmental, Resource and Ecological Economics - Paleoceanography
Environmental sciences	- Alpine Environment Data Analysis Center - Hydrology and River Basin Management - UN International Strategy for Disaster Reduction

Table 1: Research scope of GeRDI pilot communities

GeRDI is to be delivered in two modes of operation: as a set of services users can access, and as software solution communities can deploy locally. The latter option benefits from the "generic" in GeRDI, as we consider the different research practices across communities, and design with extensibility in mind.

3 Metadata as core building unit

GeRDI targets and operates on (research) metadata, not (research) data. As such, it is evident that metadata plays an important role – a central, gluing component that "reconciles" metadata harvested

[6] https://www.akademienunion.de/arbeitsgruppen/ehumanities/nfdi-arbeitsgruppe.

[7] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

[8] https://www.force11.org/group/fairgroup/fairprinciples.

across research communities and external metadata collections, and supports (metadata-based) services across communities,[9] to name the two key roles.

3.1 FAIR Metadata

Metadata in GeRDI primarily supports RD discoverability and accessibility. Every harvested research (meta)dataset, apart from its rich (bibliographic) descriptive metadata (dataset creator, title, publication date, etc.), is assigned an identifier and is added to the central GeRDI index. This corresponds to the findability principle from FAIR, as demonstrated via the GeRDI Search service. Moreover, since GeRDI adopts standardized protocols for enabling access to harvested research (meta)datasets, it also adheres to the accessibility principle from FAIR. As a side note, one GeRDI requirement is to make collected metadata FAIR. This especially comes in handy for collections which do not adhere to these principles before being harvested in GeRDI.

3.2 A common metadata structure

The great variety of research disciplines for long tail RD translates to a comparable (if not greater) metadata variety. Not all this metadata variety can be considered for harvesting or services development if an RDI is to generalize over more research communities. This leaves us with the requirement for a consistent metadata structure, i.e., a schema, for all the harvested metadata in GeRDI. A metadata schema provides a uniform/standard set of rules for information representation and ensures the consistency in metadata application/services development and its reuse. The schema will enable us capture the metadata variety from the target communities and provide a common basis to support services and internal operations.

Reconciling harvested metadata across communities implies mapping research (meta)dataset collections to central GeRDI metadata schema (cf. Section 4). Relying on this schema, GeRDI harvests long tail RD metadata and provides a variety of services beneficial to the users/research communities. The metadata elements that cannot be covered in this process due to high specificity, low priority in community use cases, or other reasons become unavailable in GeRDI services.

3.3 Metadata presence and role

In GeRDI, the first activity that involves metadata is harvesting. This activity requires a harvester for the research dataset of interest, which includes metadata mapping from the target dataset to GeRDI Schema. A harvester applies the defined mapping, which results in a dataset (metadata) description, and adds it (as a MD document) to GeRDI index. Once implemented, a Harvester control center in GeRDI manages the harvest of new and update of harvested datasets.

[9] GeRDI services range from generic, such as Search and Bookmark, to disciplinary, such as Process and Storage.

Metadata determine the services that can be supported in an RDI. Search functionality is the most prevalent in such projects. As a service, it can use any of the metadata elements that describe a dataset. On another case, once the user finds a dataset of interest, the Bookmark service enables her to create an easier reference list – in the same way we bookmark web pages of interest – and store it for future reference, via the Store service. Another interest for us is to track different interactions that users have with resources in GeRDI. Provenance aspects, such as tracking the sharing and reuse of RD or details of workflow specifications for experiments across disciplines are examples of RDI metadata-driven services.

3.4 GeRDI harvest: Some metadata stats

The last stable GeRDI release provides the following services: Harvest, Search, and Bookmark. Harvest service covers several communities and research (meta)data collections. It is worth noting that this list grows based on changing user requirements. Table 2 we provide some statistics about the harvested metadata records from GeRDI pilot communities, whereas section 6 provide more on the other services.

Community	Harvested datasets	Select subject terms/keywords
PANGAEA - Earth & Environmental Science	354424	Age, Ring width, Height above ground, Direct radiation
Zenodo	8680	Biodiversity, Taxonomy, Animalia, Arthropoda, Insecta
Esri	7991	Fitness, Sport, LCSD, Recreation
Sea Around Us - Fisheries, Ecosystems & Biodiversity	3695	Flatfishes, Species, Shelf, Temperature
European Nucleotide Archive	996	Homo sapiens, Vertebrata, Primates
German Socio-Economic Panel Study	408	Employment, Health and satisfaction indicators, Families in Germany
AlpEnDAC Project	672	Ambient, Flextra, Flexpart
OceanTEA	162	Oceanography, Conductivity, Underwater measurement
Food and Agriculture Organization of the United Nations	78	A list of states that FAO is present in
Universitätsbibliothek der Ludwig-Maximilians-Universität München	69	Medicine and Health, Architecture, History of Europe

Table 2: Research datasets per GeRDI pilot community

4 GeRDI Metadata Schema: Requirements and Approach

We shaped (and still do) the metadata schema by continuous balance (or "tension") between community requirements on one side, and RDI requirements for services support on the other. It often happens that metadata under consideration do not support a service users want to use in an RDI, and, vice-versa,

available metadata can provide more services than the ones offered by an RDI. In this section we portrait the process of GeRDI Schema development.

4.1 Metadata requirements in GeRDI

During requirements gathering with GeRDI pilot research communities, we observed that few of them adhere to a metadata standard or RD management practice, which reflects research dataset descriptions. In the context of long tail RD, this makes identifying metadata requirements for the schema more demanding.

When it comes to the metadata, we noted a variety of metadata elements, both within and across research disciplines for these communities. This variety can be organized along the following:

1. Generic metadata: Metadata present across communities that typically consists of bibliographic metadata, suitable to support RDI services that are more generic in nature. Generic metadata usually contain a smaller – but stable – number of elements.
2. Disciplinary metadata: Consist of a larger set of metadata elements, specific to research communities. For example, a research community with a scientific workflow requires a specific (RDI) service, thus corresponding metadata. This category can support disciplinary services, which are challenging to generalize.
3. Operational metadata: Enable RDI operation; users do not directly "interact" with this metadata type.

It is common for RDI projects to focus on a minimal set of elements, common across research communities – typically a generic metadata category. The problem, especially for long tail RD, is that this leaves disciplinary metadata out of the scope, and affects the level of RDI services support for the disciplines. As reported, limiting the metadata coverage in this way could seriously affect the user RDI adoption.[10] Faced with this situation, we decided to consider both metadata categories as part of the schema.

4.2 GeRDI Metadata Schema

GeRDI Schema aims to maintain a reasonable balance between generic and disciplinary metadata. In this way, we support key use cases across and within communities. This also supports potential interdisciplinary use cases enabled having multi-disciplinary RD on a single infrastructure – GeRDI. This metadata scope provides both coarse- and fine-grained metadata access, thus different use case capabilities for the research communities.

Conceptually, GeRDI Schema consists of three parts, each with a different role. Figure 1 presents GeRDI Schema components. Let's briefly present each part as per the diagram:

[10] e-IRG (2016).

- Generic part: Maps to "Generic metadata". Due to available metadata standards, we opted for reusing instead of creating our own generic schema. We chose DataCite[11] – a well-established and popular metadata standard that incentivizes RD exchange and citation (hence the "cite" in DataCite). "Generic" box in Figure 1 shows its mandatory elements.
- (Infrastructure) Extension: Maps to "Operational metadata" category. As an infrastructure, GeRDI requires certain metadata elements to support its operations, such as identify harvested resources, track resource origin, research discipline, different (URI) access links (download, view, etc.), and so on. Its role in metadata harvesting and maintenance and general RDI services support is crucial. "Extension" box in Figure 1 shows metadata elements for this part.
- Disciplinary part: Maps to "Disciplinary metadata" category, and contains metadata that are specific for research disciplines. This is the most challenging part of the schema, one of the key factors that impact both user requirements and supported services in GeRDI. "Disciplinary metadata" box in Figure 1 shows metadata organized according to the Deutsche Forschungsgemeinschaft subject areas,[12] and metadata elements provided for each area present examples of how this part of the schema could be structured. Requirements for this part are still under way.

4.3 Balancing metadata breadth and depth

Metadata (and features) that generalize over research communities (such as Search service) provide breadth, whereas metadata (and features) that specialize for given communities provide depth. In an RDI context, the former aims to serve more communities, whereas the latter aims to support individual communities better. Both with strengths and weaknesses, we need to balance between the two metadata (and feature) extremes in GeRDI Schema design.

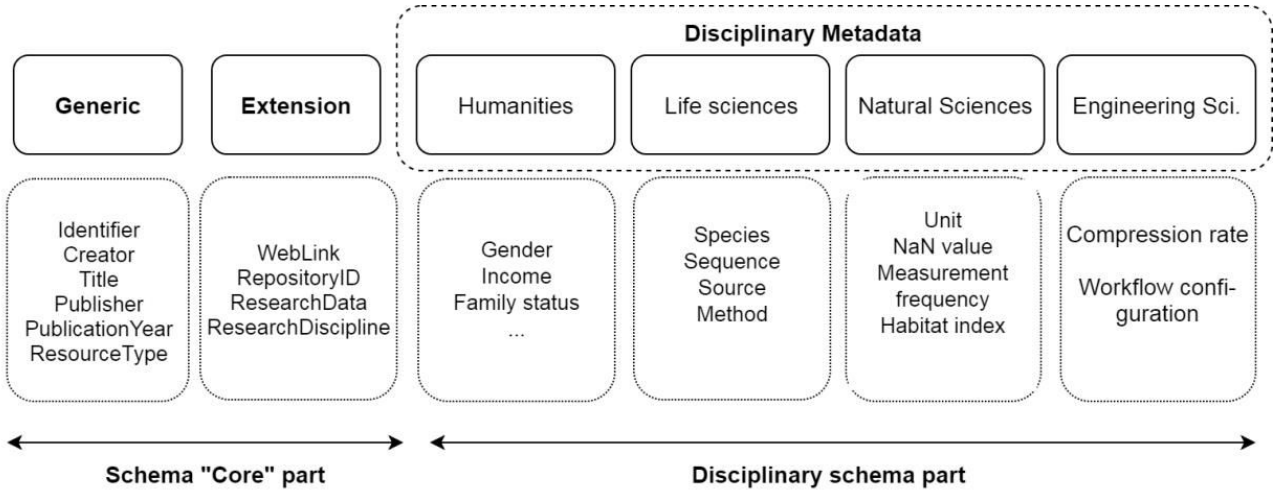

Fig. 1: GeRDI Schema components

Having this balance into view, we consider 2 approaches:

[11] http://datacite.org.

[12] http://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp.

1. Open-ended: As new use cases and new metadata elements from specific disciplines are identified, we extend GeRDI Schema accordingly (hence the term "open ended"). This produces a schema that is not finite and one that constantly changes due to (changing) requirements. Some of the advantages of this approach are that it provides maximum support for community use cases, and does not require for communities to agree on a predefined metadata set. An obvious disadvantage is its scalability: every metadata element we introduce to GeRDI Schema implies re-indexing the metadata collection to reflect schema changes.
2. Disciplinary Metadata Sets: A list of (finite) metadata elements constitutes the disciplinary part of GeRDI Schema, and does not change as new communities, new use cases, or both are considered.
   While scaling is not an issue anymore, it requires great involvement from current (and non-GeRDI) communities and does not ensure a "satisfactory" metadata set for the disciplinary part of the schema (after all, even communities from the same research discipline could perceive different metadata as important). In our attempt to finalize GeRDI Schema we are opting for this approach.

Providing a metadata standard as a backbone layer for a generic RD infrastructure project is always challenging. Having communities agree on a (minimal) priority metadata set is continuous work in progress. Starting with 9 communities across disciplines, the challenge will only increase with potential communities to join in the future. The breadth vs. depth balance is important in GeRDI as we are addressing heterogeneous (metadata) requirements.

4.4 GeRDI Services: An overview

GeRDI schema design provides a layered approach to metadata representation and access: a guaranteed, applicable-to-all services based on the "Core" part, packaged as "Core services", and a fine-grained, dedicated ones based on "disciplinary" part, packaged as "Extended services". Services combinations are also possible, whereas a discovery service, for example, based on elements from "core", could be complemented by a recommendation service based on relevant disciplinary metadata to further enrich the discovery process.

Figure 2 shows GeRDI Search service. In addition to searching GeRDI index, it has all the typical search features, including filtering based on both "core" and disciplinary schema parts. Other services in GeRDI include: Bookmarking search results for easier reference; Harvest metadata of interest and add them to GeRDI index; and Store, to download the bookmarked data (locally or on a remote storage system).

Fig. 2: GeRDI Search service

GeRDI core services – Harvest, Search, and Bookmark – are already implemented. The focus is now shifting towards extended services implementation, with Store service already under development.

5 Related work

There are many national and international RDI projects we can relate to. RADAR delivers archival and publication services for long tail RD, with inter-disciplinarity in mind, and a generic metadata schema to support its operations.[13] SowiDataNet targets small research projects from the social sciences domain that have no supporting infrastructure. Researchers can document, publish, and share RD via its portal. The metadata standard is based on Dublin Core and DataCite. The project supports resource identification and citation based on resource identifiers.[14] The international RDI project, EUDAT,[15] provides a suite of services that support the common research lifecycle. Research communities rely on its B2Find Schema, also generic in design, to describe their dataset.[16] Its B2Find service then operates on the aggregated EUDAT metadata, and supports search. Other RDI projects, smaller in scope, or focused

[13] Brophy and Razum (2017) 23; RADAR (2016).

[14] Linne and Zenk-Möltgen (2017).

[15] https://www.eudat.eu.

[16] Widmann and Thiemann (2018).

on specific disciplines also exist, that usually provide similar approaches to handling the metadata. Finally, at the EU level, the European Open Science Cloud[17] (EOSC) – via its pilot initiative, focuses on putting RDI projects under the same umbrella, experimenting with best practices and open standards across research disciplines to enable RDI interoperability.[18] Here, too, metadata plays the role of an enabler and is key to aspiring RDI projects that want to participate in EOSC.

6 Conclusion and future work

In this paper we introduce a generic RDI and propose an approach to tackle the metadata from long tail RD. GeRDI, as a distributed and federated RDI, aims to integrate research (meta)data from multiple communities and research disciplines, and provide users a "single point of view" (based on its schema) on collected metadata. In GeRDI, some services require generic and others (additional) disciplinary metadata. By reusing an existing generic standard, GeRDI adds a disciplinary metadata "layer" to complement the former. A disciplinary, lightweight metadata layer, agreed on by communities, is required to better support long tail RD communities. With this approach, GeRDI brings a metadata schema and infrastructure solution to address metadata challenges for long tail RD.

In the future we will focus on approaches for metadata enrichment and mapping to improve the information retrieval for a better support of GeRDI services. Moreover, to meet the requirements for research reproducibility, we plan to work on metadata provenance features too. Lastly, concepts alignment among disciplinary metadata will also be treated. Associating disciplinary metadata based on semantic similarities, such as "weather" and "temperature", could provide for more capable GeRDI services.

References

Union der Deutschen Akademien der Wissenschaften (2018): Der Union der deutschen Akademien der Wissenschaften zur Schaffung einer Nationalen Forschungsdateninfrastruktur (NFDI). *[Position paper]*. Available at http://www.bbaw.de/startseite-1/dateien/nfdi-positionspapier.

RfII – German Council for Scientific Information Infrastructures (2017): An International Comparison of the Development of Research Data Infrastructures. Report and Suggestions. Available at urn:nbn:de:101:1-201711084990.

Grunzke, R., Adolph, T., Biardzki, C., Bode, A., Borst, T., Bungartz, H.-J., et al. (2017): Challenges in creating a sustainable generic research data infrastructure. In: *Softwaretechnik-Trends*, 37 (2), 74–77.

[17] https://eoscpilot.eu.

[18] EOSC Pilot (2018).

e-IRG (2016): Long tail of data. e-IRG task force report 2016. Available at http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf.

Brophy, Ena, and Matthias Razum (2017): RADAR: A Research Data Management Repository for Long Tail Data. Tage 2017.

RADAR (2016). RADAR Schema 0.5. Available at https://www.radar-service.eu/en/radar-schema.

Linne, Monika, and Wolfgang Zenk-Möltgen (2017): Strengthening institutional data management and promoting data sharing in the social and economic sciences. In: *LIBER Quarterly*, 27 (1), 58–72. DOI: http://doi.org/10.18352/lq.10195.

Widmann, Heinrich and Hannes Thiemann (2018): B2FIND Integration. Eudat.eu. Available at https://www.eudat.eu/services/userdoc/b2find-integration.

EOSC Pilot (2018): About EOSCpilot. Available at https://eoscpilot.eu/about-eoscpilot.

Heidorn, P. B. (2008): Shedding Light on the Dark Data in the Long Tail of Science. In: *Library Trends*, 57 (2), 280–99.

Fidan Limani
ZBW – Leibniz Information Center for Economics
Düsternbrooker Weg 120
D-24105 Kiel
f.limani@zbw.eu

Atif Latif
ZBW – Leibniz Information Center for Economics
Düsternbrooker Weg 120
D-24105 Kiel
a.latif@zbw.eu

Timo Borst
ZBW – Leibniz Information Center for Economics
Düsternbrooker Weg 120
D-24105 Kiel
t.borst@zbw.eu

Klaus Tochtermann
ZBW – Leibniz Information Center for Economics
Düsternbrooker Weg 120
D-24105 Kiel
k.tochtermann@zbw.eu